

Limited Attention and Centrality in Social Networks

Kristina Lerman¹, Prachi Jain², Rumi Ghosh³, Jeon-Hyung Kang¹ and Ponnurangam Kumaraguru²

1. USC Information Sciences Institute, Marina del Rey, CA 90292 USA

2. Indraprastha Institute of Information Technology Delhi, India

3. HP Labs, Palo Alto, CA, USA

March 20, 2013

Abstract—How does one find important or influential people in an online social network? Researchers have proposed a variety of centrality measures to identify individuals that are, for example, often visited by a random walk, infected in an epidemic, or receive many messages from friends. Recent research suggests that a social media users’ capacity to respond to an incoming message is constrained by their finite attention, which they divide over all incoming information, i.e., information sent by users they follow. We propose a new measure of centrality — limited-attention version of Bonacich’s Alpha-centrality — that models the effect of limited attention on epidemic diffusion. The new measure describes a process in which nodes broadcast messages to their out-neighbors, but the neighbors’ ability to receive the message depends on the number of in-neighbors they have. We evaluate the proposed measure on real-world online social networks and show that it can better reproduce an empirical influence ranking of users than other popular centrality measures.

I. Introduction

An individual’s position within a social network is thought to confer advantages, allowing him to exploit the structure of social ties to accumulate power, prestige or influence [3], [18], [11], [25], [7], [8]. Many measures of centrality were proposed to capture the importance of the position in a network. Some of these, like degree and betweenness centrality [11], measure an individual’s ability to control the flow of information in the network. Other measures give higher centrality to those positions that are themselves connected to central positions [23], [4], [29], [5]. The growing popularity of online social media has sparked new interest in centrality. Researchers have proposed using centrality to identify influential social media users [9], [2] whose endorsement can, for example, maximize the reach of a “viral” marketing campaign [24], or conversely, who can most quickly stop a malicious rumor from spreading.

Most of the existing centrality measures examine link structure of the network to identify key nodes within it. Take, for example, the Web, which is represented as a directed graph of hyperlinked Web pages. An important page within this graph is one that is visited often by Web surfers. This observation forms the basis of Google’s

original Web page ranking algorithm PageRank [29]. By modeling Web surfing as a random walk, PageRank assigns a centrality score to each page based on its value in the equilibrium distribution of the random walk. However, a central individual in a social network through which disease is spreading is one who infects, either directly or indirectly, most others. Unlike Web surfing, the spread of a virus is modeled as an epidemic process. Thus, PageRank, which is intimately connected with random walks, will not identify key individuals in a social network. Instead, a measure such as the Katz score [23] or Bonacich’s Alpha centrality [4], which gives the equilibrium distribution of an epidemic process on a network [14], is more appropriate.

Now consider information spreading through an online social network, for instance, by users sending messages or product recommendations to their friends. While information spread in networks is often modeled as an epidemic process (e.g., [19], [28]), recent research suggests that psychological and cognitive factors are important in determining whether a person will *see* and *act* on friends’ recommendations. Specifically, attention was shown to be a critical aspect of online behavior [17], [33], [32], [20]. Attention is the psychological mechanism that controls how we process incoming stimuli and decide what activities to engage in [22], [30]. Actions, such as reading a tweet, browsing a Web page, or responding to email, require mental effort, and since human brain’s capacity for mental effort is limited, so is attention. Moreover, online users must divide their attention over all incoming stimuli [20]. As a consequence, the more stimuli people have to process, the smaller the probability they will respond to any one stimulus. While attention need not be distributed uniformly over friends — some friends may receive a greater share of a person’s attention due to familiarity, trust, social closeness, or influence [16], [21] — for simplicity, we assume that each friend receives the same fraction of a person’s attention. We call this phenomenon *limited attention* (*la*).

Limited, divided attention changes the nature of interactions between nodes in a network and therefore, how central nodes are identified. Now a node’s capacity to

infect others depends not only on how many connections it has but also on who and how many others these nodes are connected to. In Section III, we introduce a new centrality measure — limited-attention Alpha-Centrality (*laAC*) — that models attention-limited nature of social interactions and provide its mathematical definition. For completeness, we also introduce and define limited-attention PageRank (*laPR*), which models the effect of limited attention on a random walk process. In Section IV, we evaluate the proposed algorithms and centrality measures on real-world data, including follower graphs from social media sites Digg and Twitter. In the Appendix, we present fast approximate algorithms that allow us to calculate these measures even on large graphs and provide their performance guarantees.

II. Dynamics, Attention and Centrality

Centrality measures examine topology of a network to identify important or central nodes within it. It has been recognized recently, however, that centrality is the product of a network's *links* and the *dynamical processes* taking place on it, which determine how ideas, pathogens, or influence flow along social links [6], [26], [15], [14]. Take, for example, one definition of centrality used by the popular PageRank algorithm [29]: a network node is important if it is often visited by a random walk. A random walk is a stochastic process that starts at some node, and at each time step transitions to a randomly selected neighbor of the current node. Variants of the random walk are used to model flows in physical systems, e.g., chemical and heat diffusion, and can be used to model social phenomena resulting from one-to-one interactions, such as Web surfing, money exchange and phone conversations.

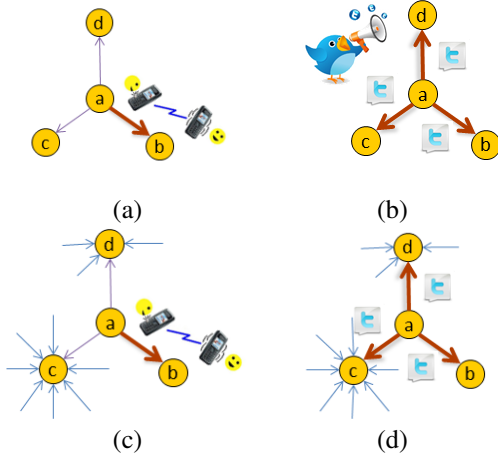


Fig. 1. Different dynamical processes taking place on a network: (a) random walk, (b) epidemic spread, and their limited-attention variants: (c) limited-attention random walk and (d) limited-attention epidemic spread. In limited-attention process, a node's capacity to receive a message depends on its in-degree.

In a social network, a message or a virus propagates by being broadcast by an infected individual to *all* her (out-) neighbors. Such processes are modeled as an epidemic

(or a contact) process. The difference between it and the random walk is illustrated in Figure 1, which shows the neighborhood of node *a*. Directed edges in this network represent, for example, hyperlinks between Web pages, or who can call whom in a social network, or in the context of social media, they can also indicate that *b*, *c* and *d* follow *a* and receive broadcasts from her. Figure 1(a) illustrates a one-to-one interaction, e.g., phone call, while Fig. 1(b) shows a one-to-many broadcast.

Until now, we have assumed that nodes have an unlimited capacity to receive incoming signals, whether Web surfers, phone calls, or messages from friends. This may not always be the case. Suppose a Web server can receive a limited number of connections, in extreme case only one. Then the probability that a Web surfer starting at *a* will reach *b* depends on whether the Web server in charge of *b* is able to receive an incoming request. In a social network, cognitive and perceptual factors can limit a person's capacity to process incoming messages [20]. Such factors collectively figure into the phenomenon we refer to as *limited attention*. This means that the probability a user will respond to a message from a friend decreases with the number of friends she follows. This is illustrated graphically in Fig. 1(c) and (d). Node *b* is more likely to receive a message from *a* than node *c* because *c* is receiving messages from eight nodes, while *b* from only one node.

Different dynamic processes lead to different notions of centrality. PageRank is used to find nodes that are often visited by a random walk (with random restarts), while Alpha- (or Bonacich) Centrality identifies nodes that are often infected during an epidemic [14]. Below we define limited-attention PageRank and limited-attention Alpha-Centrality, centrality measures that take into account the finite attention of online social users. Limited-attention PageRank identifies nodes that are often visited by a random walk, when each node's capacity to receive the walker depends on its in-degree. Similarly, limited-attention Alpha-Centrality identifies nodes that are often infected in an epidemic, when each node's susceptibility to infection also depends on its in-degree.

III. Limited-Attention Centrality

We represent a network as a directed graph with V nodes and E edges. The adjacency matrix of the graph is defined as: $A[u, v] = 1$ if there is an edge from u to v ; otherwise, $A[u, v] = 0$. Also, $A[u, u] = 0$. The set of out-neighbors of u is $\{v \in V | (u, v) \in E\}$; and the set of in-neighbors is $\{v \in V | (v, u) \in E\}$. Two other important quantities are the in-degree and out-degree matrices. The out-degree matrix D_{out} is a diagonal matrix defined as $D_{out}[i, i] = \sum_j A[i, j] = Ae^T$ and $D_{out}[i, j] = 0 \forall i \neq j$. Here, e is a $|V|$ -dimensional row vector of ones, and e^T is its transpose. The in-degree matrix D_{in} is a diagonal matrix defined as $D_{in}[i, i] = \sum_i A[i, j] = eA$ and $D_{in}[i, j] = 0 \forall i \neq j$.

A. Limited-attention PageRank: A PageRank vector $\text{pr}(\alpha, s)$ is the steady state probability distribution of a random walk with restarts with a damping factor α . This means that with a probability α , the walk transitions to one of the out-neighbors of a current node, and with probability $(1-\alpha)$ it transitions to any node in the network. The starting vector s , gives the probability distribution for where the walk transitions after restarting, which is usually taken as a uniform vector $s = e/|V|$. The transfer matrix $D_{out}^{-1}A$ encodes the transition probabilities of a random walk on the network. PageRank vector $\text{pr}(\alpha, s)$ is the unique solution of the following iterative equation:

$$\text{pr}(\alpha, s) = (1 - \alpha)s + \alpha \text{pr}(\alpha, s) D_{out}^{-1}A \quad (3.1)$$

Now, if a node's capacity to receive a random walker is limited, the transfer matrix must be modified. As stated above, we consider the simplest scenario in which the finite capacity is divided uniformly between all incoming connections. This case is modeled by the transfer matrix $D_{out}^{-1}AD_{in}^{-1}$. Therefore, limited-attention PageRank $^{la}\text{pr}(\alpha, s)$ is the solution of the following iterative equation:

$$^{la}\text{pr}(\alpha, s) = (1 - \alpha)s + \alpha ^{la}\text{pr}(\alpha, s) D_{out}^{-1}AD_{in}^{-1} \quad (3.2)$$

The starting vector above is $s = eD_{in}^{-1}$. Note that while the PageRank transfer matrix $D_{out}^{-1}A$ is stochastic, since each row or column sums to one, this is no longer the case for the limited-attention PageRank transfer matrix.

We illustrate the differences between PageRank and limited-attention PageRank on a toy directed network. Figure 2(a) shows this network with the size of the node proportional to its centrality score relative to other nodes, as determined by PageRank (with $\alpha = 0.85$). Node B is the most central, since it has many in-links, enabling a random walker to reach it via many different paths. Peripheral nodes H, I, J , etc., are less important, since they only receive the random walker via a random jump. On the other hand, limited-attention PageRank, shown in Fig. 2(b), scores these nodes highly. The node ranked highest by PageRank, B , on the other hand, dramatically decreases in centrality. This node divides its attention among many in-links, limiting its ability to receive a random walker along any specific link. The peripheral nodes, on the other hand, have few in-links, and are better able to receive the random walker, whether it is following an out-link or executing a random jump. Their importance, therefore, is greater in this scenario.

B. Limited-attention Alpha-Centrality: Alpha-Centrality measures the total number of paths from a node, exponentially attenuated by their length. Bonacich introduced this measure [4] as a generalization of the index of status proposed by Katz [23], and it is sometimes referred to as Bonacich centrality. Alpha-Centrality matrix gives the number of attenuated paths between two nodes, and it is usually written as a power series expansion of the adjacency matrix, with attenuation parameter $\alpha \geq 0$:

$C = A + \alpha A^2 + \alpha^2 A^3 + \alpha^3 A^4 + \dots$. This series converges to $C = \alpha A(I - \alpha A)^{-1}$ while $\alpha < 1/\lambda_{max}$, where λ_{max} is the largest eigenvalue of A (i.e., spectral radius of the network). Parameter α determines how far, on average, a node's effect will be felt and sets the length scale of interactions. When α is small, Alpha-Centrality probes only the local structure of the network. As α grows, more distant nodes contribute to the centrality score of a given node [13]. As $\alpha \rightarrow 1/\lambda_{max}$, the length scale of interactions diverges and it becomes a global measure.

Alpha-Centrality gives the steady state distribution of an epidemic process on a network [14], where α is the probability to transmit a message or influence along a link. Therefore, (i, j) th entry of the Alpha-Centrality matrix C can be interpreted as the likelihood that the virus will reach node j from node i . Summing over all columns j gives the Alpha-Centrality score of node i , $\text{ac}(\alpha) = Ce^T = \sum_j C(i, j)$, or the number of infections directly or indirectly caused by node i . Summing over the rows of the Alpha-Centrality matrix, on the other hand, gives $\text{ac}(\alpha)^T = e \cdot C = \sum_i C(i, j)$, the total number of times that node i is infected by others.

Alpha-Centrality vector $\text{ac}(\alpha, s)$ can also be defined iteratively as:

$$\text{ac}(\alpha, s) = s + \alpha A \cdot \text{ac}(\alpha, s), \quad (3.3)$$

where the starting vector $s = Ae^T$ is taken as out-degree centrality [4].

Let us now consider the case in which a node's capacity to receive incoming stimuli — whether messages or viruses — is limited and uniformly divided among all incoming connections. Therefore, the probability that node j will receive a message broadcast by i will be proportional to $1/d_{in}(j)$, where $d_{in}(j)$ is the in-degree of node j . The limited-attention Alpha-Centrality matrix can be written in terms of the modified adjacency matrix $M = AD_{in}^{-1}$ as:

$$C_{la} = M + \alpha M^2 + \alpha^2 M^3 + \alpha^3 M^4 + \dots$$

The limited-attention Alpha-Centrality vector $^{la}\text{ac}(\alpha, s)$ can also be written in iterative form:

$$^{la}\text{ac}(\alpha, s) = s + \alpha AD_{in}^{-1} \cdot ^{la}\text{ac}(\alpha, s), \quad (3.4)$$

with the starting vector $s = AD_{in}^{-1}e^T$. Note that the transfer matrix AD_{in}^{-1} is a stochastic matrix.

Figures 2(c) and (d) illustrate the differences between Alpha-Centrality and its limited-attention variant. Figure 2(c) shows the directed network with nodes sizes proportional to their ac scores. The Alpha-Centrality scores in this example were calculated for $\alpha = 0.85$. The rankings of nodes are similar to those produced by PageRank (Fig. 2(a)), though node E , for example, is relatively less important. In the limited-attention variant, shown in Fig. 2(d), the picture looks completely different. While B in (d) loses its importance, due to many in-links, node A becomes more central, since it receives incoming signals over a single in-link. Peripheral nodes are not judged to be central, because, unlike random jumps in PageRank, they never receive any signals.

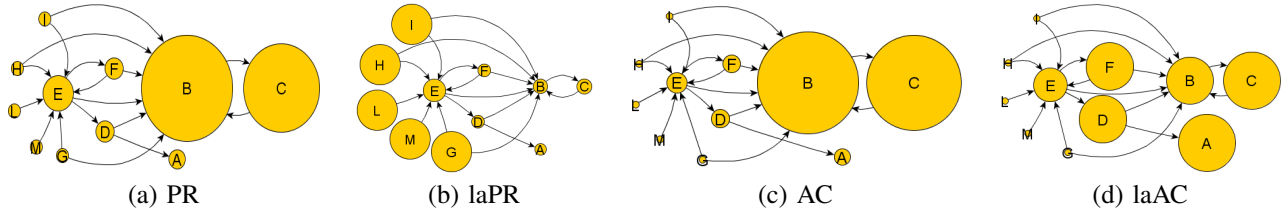


Fig. 2. Directed network with sizes of nodes weighed by their score according to (a) PageRank and (b) attention-limited PageRank (c) Alpha-centrality and (d) limited-attention Alpha-centrality of the influence graph.

IV. Applications to Social Media

We use centrality measures proposed in this paper to identify influential people on social media. Correctly identifying such people can have far-reaching consequences for identifying noteworthy content, targeted information diffusion, and other applications. While calculating Eq. 3.4 was infeasible for such large networks, we used approximate algorithms presented in the Appendix for these calculation. Appendix also gives performance guarantees of the approximate algorithms.

Researchers have proposed a number of simple heuristics to identify influential social media users that rely, for example, on the number of followers or mentions [9], [27], [2]. Others have used centrality by analyzing the follower graph to find users with high PageRank scores [10], [31]. However, since information spread on networks is traditionally described as an epidemic [19], [28], Alpha-Centrality may do a better job [12], since it explicitly models epidemic dynamics. We show, however, that limited-attention Alpha-Centrality, the measure that accounts for both the epidemic nature of social media broadcasts and the divided attention of its users, does a better job identifying influential users than Alpha-Centrality.

Specifically, we study URL-sharing activity on Digg and Twitter, two popular social media sites for content sharing. Both sites allow users to follow other users by listing them as friends. The follower relation is asymmetric. When user A follows (becomes a fan of) B , she receives B 's broadcasts, but not vice versa: we denote the relationship as $B \rightarrow A$. Representing the follower graph in matrix form, a user's out-degree measures the number of followers she has, and her in-degree the number of friends she follows.

A. Data Collection: The Digg dataset contains more than 3 million votes on some 3500 stories promoted to Digg's front page in June 2009. More than 139K distinct users voted for at least one story in the data set (submission counts as the story's first vote). We call these users *active* users. Next, we extracted the friendship links created by active users and constructed a follower graph that contained active users who were following the activities of others. Only about 71K active users listed others as friends, resulting in network with around 280K users and over 1.7 million links.

The Twitter data set was collected over a period

of three weeks in October 2010 using the Gardenhose streaming API. We focused on tweets that included a URL in the body of the message. In order to ensure that we had the complete tweeting history of the URL, we used the search API to retrieve all tweets containing that URL. Users who tweeted the URL are considered *active*. Data collection process resulted in more than 3 million posts tweeted by 816K users which mentioned 70K distinct URLs. Next, we used the REST API to collect followers of each active user, keeping only those followers who themselves were active, i.e., tweeted at least one URL during data collection period. The resulting follower graph had almost 700K nodes and over 36 million edges. More details of the data collection method are provided in [14].

B. Results: We calculate Alpha-Centrality (AC) and limited-attention Alpha-Centrality ($laAC$) on the Digg and Twitter follower graphs using algorithm for $laAC$ (Alg. 2) presented in the Appendix and the algorithm for AC presented in [15]. These are approximate algorithms with proven performance guarantees. We calculate limited-attention PageRank ($laPR$) on the transpose of the follower graph using Alg. 1, since node's influence is related to the number of walks it generates, rather than receives. The in- and out-degrees were conditioned by adding a small number (0.01) to avoid division to zero.

In order to compare the performance of centrality measures, we need a relevant measure of influence. When a user posts a URL on Digg or Twitter, she broadcasts it to all her followers. We refer to this user as the *submitter*. Whether or not her follower will re-broadcast the URL (i.e., retweet it on Twitter or vote for it on Digg) depends on its *quality* and *submitter's influence*. Assuming that URL's quality is uncorrelated with the submitter, we can average out its effect by aggregating over all URLs submitted by the same user [12]. The residual difference between submitters can be attributed to variations in influence. Similar to [9], [14], [2], we use the average number of times the URLs submitted by the user are re-broadcast by her followers as the *empirical measure of influence*.

Figure 3 shows how well the rankings produced by different centralities correlate with the empirical influence rankings of users who submitted at least two URLs which were rebroadcast at least ten times. We use Spearman rank correlation because it is less sensitive to variations in scores, and we expect some variation to arise in approxi-

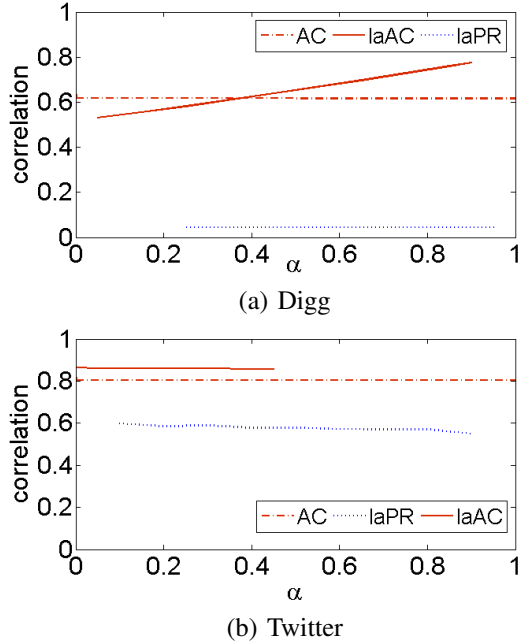


Fig. 3. Correlation of rankings of (a) Digg and (b) Twitter users found by different measures of centrality with the empirical influence ranking.

mate centrality scores. Limited-attention Alpha-Centrality correlates better with the empirical measure of influence than Alpha-Centrality over a broad range of α values, consistent with our claim that $laAC$ is a better measure for predicting central social media users, because it better models the dynamics of online communication than AC . On Digg, AC appears to outperform $laAC$ for small values of α . Since α can be thought of as the scale of interaction, this implies that locally, AC better predicts influential users. This could be the consequence of the fact that our measure of influence, i.e., number of re-broadcasts by followers, is a local measure. In the future, we plan to compare the performance of centrality measures using a global measure of influence, for example, the average size of cascades triggered by submitted URLs. We did not expect limited-attention PageRank ($laPR$) to predict influence rankings of Digg and Twitter users, since the dynamic process this centrality models does not at all describe communication patterns of social media users, and we found no correlation.

Interestingly, PageRank and $laAC$ have similar performance, since $laAC$ calculated on the adjacency matrix A of the follower graph is almost identical to PR calculated on the transpose of A , except that the starting vectors are different in the two algorithms. This suggests that dynamics of random walk are almost equivalent to epidemic dynamics under the conditions of uniformly divided attention, when direction of the flow is reversed. This observation could explain why PR can give good results in the social media domain. We leave implications of this observation for future research.

V. Conclusion

Information flow in social networks, including online networks, is often modeled as an epidemic process, suggesting that centrality measures based on epidemics are appropriate for predicting influential social media users. We propose a new centrality measure that takes into account the finite capacity of social media users to process incoming messages from friends. We modeled such limited attention by scaling the probability a node receives a message by the inverse of its in-degree. We presented approximate algorithm that allows us to efficiently calculate proposed measure for the real-world social networks on Digg and Twitter. We showed empirically that centrality measure that models limited-attention epidemics does a better job predicting highly retweeted social media users than one that models simple epidemics. Our findings suggest that the nature of interactions among network nodes should determine how central nodes are identified.

Acknowledgements: This material is based upon work supported by the Air Force Office of Scientific Research under contracts FA9550-10-1-0569 and FA9550-10-1-0102, by the Air Force Research Laboratories under contract FA8750-12-2-0186, by DARPA under contract W911NF-12-1-0034, and by the National Science Foundation under grant CIF-1217605. PJ's internship was sponsored by the USC Viterbi-India Summer program.

Appendix: Approximate Algorithms

Finding limited-attention PageRank (Eq. 3.2) and Alpha-Centrality (Eq. 3.4) requires the computation of matrix inverse, which can be done in $O(|V|^3)$ operations using the naive implementation of the algorithm ($|V|$ is the number of nodes in the network). This is prohibitively expensive for networks with thousands or more nodes. However, solving equations iteratively requires $O(|V|^2)$ operations in each iteration, though we do not know how many iterations are sufficient for an optimal solution. We propose Approximate Limited-Attention Page Rank and Approximate Limited-Attention Alpha Centrality algorithms, which can be used to calculate a near optimal solution. The algorithms use a single error tolerance parameter δ ($0 < \delta \leq 1$) to control both the quality of the solution and computation time.

The proposed algorithms and their performance guarantee are based on the approximate PageRank [1] and approximate Alpha-Centrality [15] algorithms. They provide a flexible way to compute the near optimal centrality vector \tilde{c} using a starting vector s and a residual vector r . Initially $r = s$ and $\tilde{c} = \vec{0}$. The algorithms iteratively move the weight from r to \tilde{c} vector, until the values in the residual vector r are sufficiently small. The amount of error in the approximate centrality vector is equivalent to the amount remaining in the residual vector. The performance guarantee of the proposed algorithms are given in Theorem 0.1 and Theorem 0.2, which are based on Lemma 0.1. The Lemma states that each iteration maintains an invariant

vector $\tilde{c}r = cr(s) - cr(r) = cr(s - r)$. This means that the amount of error in the approximate centrality vector is equivalent to the error remaining in the residual vector.

PROPOSITION 0.1. *For any fixed value of α in $[0, 1]$ and starting vector s , $cr(\alpha, s)$ is linear in s .*

Proof: The limited-attention PageRank vector $^{la}pr(\alpha, s)$ is a unique solution to

$$cr(s) = ^{la}pr(\alpha, s) = (1 - \alpha)s + \alpha \cdot ^{la}pr(\alpha, s)M$$

where $M = D_{out}^{-1}AD_{in}^{-1}$. The limited-attention Alpha-Centrality vector $^{la}ac(\alpha, s)$ can also be written in iterative form:

$$cr(s) = ^{la}ac(\alpha, s) = s + \alpha \cdot ^{la}ac(\alpha, s)M,$$

where $M = AD_{in}^{-1}$. The centrality vectors can be proved linear with respect to s by substituting suitable values for $cr(s)$ and M in the proof presented in [15]. ■

LEMMA 0.1. *At the start of each iteration of while loop $\tilde{c}r = cr(s) - cr(r) = cr(s - r)$ such as the sum of elements in r decreases with each iteration.*

Proof: The proof of correctness is based on Proposition 0.1. During initialization, $r = s$ and $\tilde{c}r = \vec{0}$; therefore, $cr(s - r) = cr(\vec{0}) = \vec{0} = \tilde{c}r$. The lemma is maintained throughout the execution of the loop. To prove this, we use a row vector z_u such as $z_u(i) = 1$ if $i = u$; otherwise, $z_u(i) = 0$. Before the next iteration of while loop in Algorithm 1 we have $\tilde{c}r = \tilde{c}r + (1 - \alpha)z_i r(i)$ and $r' = r - z_i r(i) + \alpha r(i)Mz'$ where $\tilde{c}r', r'$ are updated centrality vector and residual vectors and i is the vertex dequeued in line number 11 of the algorithm. Now consider

$$\begin{aligned} cr(r) &= cr(r - z_i r(i)) + cr(z_i r(i)) \\ &= cr(r - z_i r(i)) + (1 - \alpha)z_i r(i) + cr(\alpha z_i r(i)M) \\ &= cr(r - z_i r(i) + \alpha z_i r(i)M) + (1 - \alpha)z_i r(i) \\ &= cr(r') + \tilde{c}r' - \tilde{c}r = cr(r') + \tilde{c}r' - cr(s - r) \end{aligned}$$

It follows that $\tilde{c}r' = cr(r) - cr(r') + cr(s - r) = cr(r - r' + (s - r)) = cr(s - r')$. On termination of the loop, given the lemma and an error tolerance parameter the approximate centrality vector should always satisfy

$$cr(s)[i] \geq \tilde{c}r[i] \geq (1 - \delta)cr(s)[i] \quad \forall i \in V$$

We choose a uniform starting vector s , $s[i] = \|s\|_1/|V|$, $\forall i \in V$. The algorithm terminates when $r[i] \leq \epsilon d_{out}^{max}$; $\forall i \in V$, so we choose $\epsilon = \frac{\delta \|s\|_1}{|V|d_{out}^{max}} = \frac{\delta s[i]}{d_{out}^{max}}$. With this choice of ϵ we also ensure freedom in choice of the value of α with in the range of 0 to 1. This freedom is achieved at the cost of increased running time of the algorithm. In the end $r[i] \leq \delta s[i]$, therefore, $\implies cr(r)[i] \leq \delta cr(s)[i]$. Thus,

$$\tilde{c}r[i] \geq (1 - \delta)cr(s)[i].$$

It is obvious that $cr(s)[i] \geq \tilde{c}r[i]$; hence $cr(s)[i] \geq \tilde{c}r[i] \geq (1 - \delta)cr(s)[i] \quad \forall i \in V$. Also the sum of all elements of

residual vector $\sum r'$ is

$$\sum r' = \sum r - r[i] + \left(\alpha \frac{r[i]}{d_{out}(i)} \cdot \sum_{j \in N^{out}(i)} \frac{1}{d_{in}(j)} \right)$$

Since value of α lies in $[0, 1]$ and $\sum_{j \in N^{out}(i)} \frac{1}{d_{in}(j)} \leq d_{out}(i)$, net sum of all values of residual vector decreases with each iteration of while loop. Similarly the we can prove that the lemma is valid for Algorithm 2. ■

A. Approximate Limited-Attention PageRank: Limited attention Page Rank ($laPR$) given by Eq. 3.2, can be written as the solution $cr(\alpha, s)$ of:

$$cr(\alpha, s)[j] = (1 - \alpha)s[j] + \alpha \sum_{i \in N^{in}(j)} \frac{cr(\alpha, s)[i]}{d_{out}(i)d_{in}(j)}.$$

Here $N^{in}(j)$ is a set of in-neighbors of j , i.e., nodes i such that edge $(i, j) \in E$. Also, $N^{out}(j)$ is the set of out-neighbors of j , i.e., nodes i such that $(j, i) \in E$. We take the starting vector $s = e/|V|$ to be uniform. To simplify notation, we will refer to $cr(\alpha, s)$ as cr .

Algorithm 1 Approximate limited-attention PageRank(V, E, s, α, δ)

```

1:  $\epsilon = \delta \|s\|_1 / |V| d_{out}^{max}$ ;
2:  $r = s$ ;
3: Queue  $q = \text{new Queue}()$ ;
4: for each  $i \in V$  do
5:    $\tilde{c}r[i] = 0$ ;
6:   if  $\frac{r[i]}{d_{out}^{max}} > \epsilon$  then
7:      $q.add(i)$ ;
8:   end if
9: end for
10: while  $q.size() > 0$  do
11:    $i = q.dequeue()$ ;
12:    $\tilde{c}r[i] = \tilde{c}r[i] + (1 - \alpha)r[i]$ ;
13:    $T = \alpha r[i] / d_{out}(i)$ ;
14:    $r[i] = 0$ ;
15:   for each  $j \in N^{out}(i)$  do
16:      $r[j] = r[j] + T / d_{in}(j)$ ;
17:     if  $!q.contains(j)$  and  $r[j] / d_{out}^{max} > \epsilon$  then
18:        $q.add(j)$ ;
19:     end if
20:   end for
21: end while
22: return  $\tilde{c}r$ ;

```

THEOREM 0.1. *Given an $0 \leq \alpha < 1$ and a uniform starting vector s , the approximate centrality vector $\tilde{c}r$ is obtained from the algorithm in run time $O\left(\frac{|V|d_{out}^{max}}{(1-\alpha)\delta}\right)$.*

Proof: Given an α in $[0, 1]$. Algorithm 1 works by dividing $\alpha r[i]$ equally amongst all $N^{out}(i)$ out-neighbors of node i . Each out-neighbor j receives a fraction of the weight, based on its capacity, $d_{in}(j)$, to receive incoming

messages. Hence, all $r[j]$ will increase by some fraction. Let r be old residual vector and r' be the updated residual vector. The sum of all elements of residual vector $\sum r'$ is

$$\sum r' = \sum r - r[i] + \left(\alpha \frac{r[i]}{d_{out}(i)} \cdot \sum_{j \in N_{out}(i)} \frac{1}{d_{in}(j)} \right)$$

The sum of the entries of residual vector decreases by

$$\begin{aligned} \sum r &= \left(\sum r - r[i] + \frac{\alpha r[i]}{d_{out}(i)} \sum_{j \in N_{out}(i)} \frac{1}{d_{in}(j)} \right) \\ &= r[i] - \left(\alpha \frac{r[i]}{d_{out}(i)} \cdot \sum_{j \in N_{out}(i)} \frac{1}{d_{in}(j)} \right) \\ &> r[i] - \left(\alpha \frac{r[i]}{d_{out}(i)} \cdot d_{out}(i) \right) > (1 - \alpha) \epsilon d_{out}^{max} \end{aligned}$$

Let k be the total number of iterations, net amount removed from residual vector will be at least

$$k(1 - \alpha) \epsilon d_{out}^{max} < \|s\|_1 \implies k < \frac{\|s\|_1}{(1 - \alpha) \epsilon d_{out}^{max}}$$

Since each iteration is proportional to $d_{out}[i]$, the worst case time complexity is $O(\frac{\|s\|_1}{(1 - \alpha) \epsilon})$. For our choice of ϵ , this is equivalent to $O(\frac{|V| d_{out}^{max}}{\delta(1 - \alpha)})$. ■

B. Approximate Limited-Attention Alpha-Centrality: Limited attention Alpha-Centrality (*laAC*), given by Eq. 3.4, can be rewritten as the solution $cr(\alpha, s)$ of:

$$cr(\alpha, s)[i] = s[i] + \alpha \sum_{j \in N_{out}(i)} \frac{cr(\alpha, s)[j]}{d_{in}(j)},$$

with the starting vector $s[i] = \sum_{j \in N_{out}(i)} 1/d_{in}(j)$. As before, we use $N_{out}(i)$ to denote the set of out-neighbors, and $N_{in}(i)$ the in-neighbors, of node i .

Algorithm 2 Approximate Limited-Attention Alpha-Centrality(V, E, s, α, δ)

```

1:  $\epsilon = \frac{\delta \|s\|_1}{|V| d_{in}^{max}};$ 
2:  $r = s;$ 
3: Queue  $q = \text{new Queue}();$ 
4: for each  $i \in V$  do
5:    $\tilde{cr}[i] = 0;$ 
6:   if  $\frac{r[i]}{d_{in}^{max}} > \epsilon$  then
7:      $q.add(i);$ 
8:   end if
9: end for
10: while  $q.size() > 0$  do
11:    $i = q.dequeue();$ 
12:    $\tilde{cr}[i] = \tilde{cr}[i] + r[i];$ 
13:    $T = \alpha \cdot \frac{r[i]}{d_{in}(i)};$ 
14:    $r(u) = 0;$ 
15:   for each  $j \in N_{in}(i)$  do
16:      $r[j] = r[j] + T;$ 
17:     if  $!q.contains(i)$  and  $\frac{r[j]}{d_{in}^{max}} > \epsilon$  then
18:        $q.add(j);$ 
19:     end if
20:   end for
21: end while
22: return  $\tilde{cr};$ 

```

THEOREM 0.2. Given $0 < \alpha < 1$ and starting vector s , the approximate centrality vector \tilde{cr} is obtained from the algorithm in run time $O(\frac{|V| d_{in}^{max}}{\delta(1 - \alpha)})$.

Proof: Given an α in $[0, 1]$. Let r be old residual vector and r' be the updated residual vector. The sum of all elements of residual vector $\sum r'$ is

$$\begin{aligned} \sum r' &= \sum r - r[i] + \left(\alpha d_{in}(j) \cdot \frac{r[j]}{d_{in}(j)} \right) \\ &= \sum r - r[j] + (\alpha r[j]) \end{aligned}$$

The sum of the entries of residual vector decreases by

$$\begin{aligned} \sum r &= \left(\sum r - r[j] + (\alpha r[j]) \right) \\ &= r[j] - (\alpha r[j]) \\ &= (1 - \alpha) r[j] > (1 - \alpha) \epsilon d_{in}^{max} \end{aligned}$$

Let k be the total number of iterations, net amount removed from residual vector will be at least

$$k(1 - \alpha) \epsilon d_{in}^{max} < \|s\|_1 \implies k < \frac{\|s\|_1}{(1 - \alpha) \epsilon d_{in}^{max}}$$

Since each iteration is proportional to d_{in} , so the worst case time complexity is $O(\frac{\|s\|_1}{(1 - \alpha) \epsilon})$. For our choice of ϵ this is equivalent to $O(\frac{|V| d_{in}^{max}}{\delta(1 - \alpha)})$. ■

C. Performance of Approximate Algorithms: For relatively small networks (up to thousands of nodes), we compared centrality scores calculated by the approximate algorithms to those calculated by their exact versions.

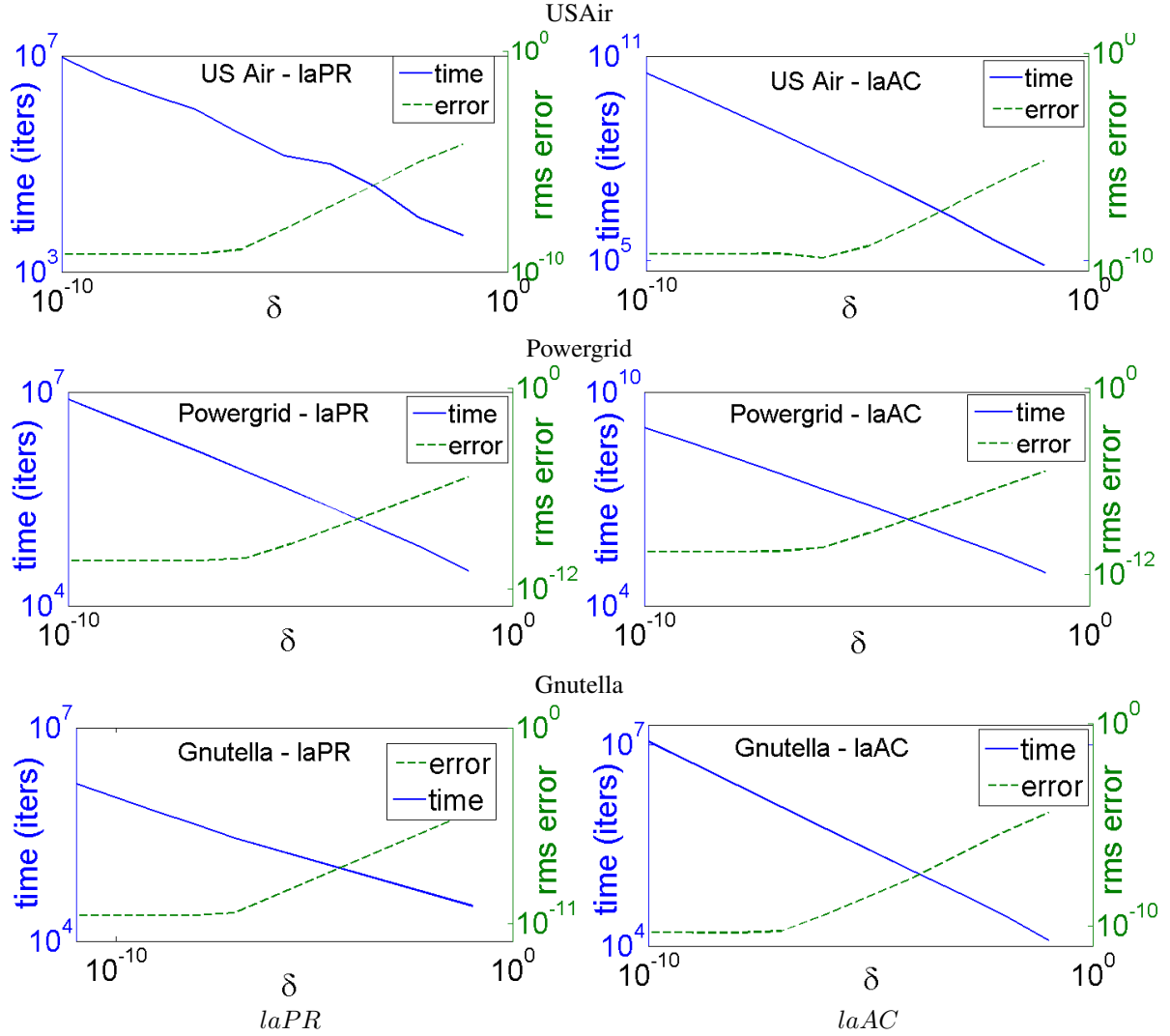


Fig. 4. Performance of the fast approximate limited-attention PageRank (*laPR*) and Alpha-Centrality (*laAC*) on Gnutella, US Air and Power grid networks. Performance is measured by time (number of iterations of the approximate algorithm) and *rms* error of the centrality values calculated by the approximate and exact algorithms.

The *USAir* network¹ is an undirected network of 332 nodes and 4,252 edges, which represent airports linked by direct flights. The *Powergrid* network² is an undirected network of 4,941 nodes and 6,594 edges representing the topology of the US Western States power grid. The Gnutella dataset³ contains a snapshot of the Gnutella peer to peer network with 6,301 nodes and 20,777 edges.

Figure 4 shows the performance of the fast approximate algorithms proposed in this paper on the three networks vs the error tolerance δ . Performance is measured in terms of time (number of iterations) taken to compute approximate centrality values and *rms* error of these compared to the values computed by the exact algorithms

Eqs. 3.2 and 3.4. In all cases, while it takes longer to compute centrality scores for decreasing values of δ , the answers are closer to their exact values.

Figure 5 plots the number of iterations taken by the proposed algorithms to calculate centralities for the Digg and Twitter data sets for different values of the error tolerance parameter δ . Parameter values used in the calculations were $\alpha = 9.0 \times 10^{-4}$ for both *laAC* and *laPR* on Digg, and $\alpha = 1 \times 10^{-4}$ for *laAC* and $\alpha = 0.9$ for *laPR* on Twitter. As expected, the number of iterations increases for smaller error tolerances.

References

- [1] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Proc IEEE Foundations of Computer Science*, pages 475–486, 2006.
- [2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts.

¹<http://vlado.fmf.uni-lj.si/pub/networks/data/>

²<http://cdg.columbia.edu/cdg/datasets>

³<http://snap.stanford.edu/data/>

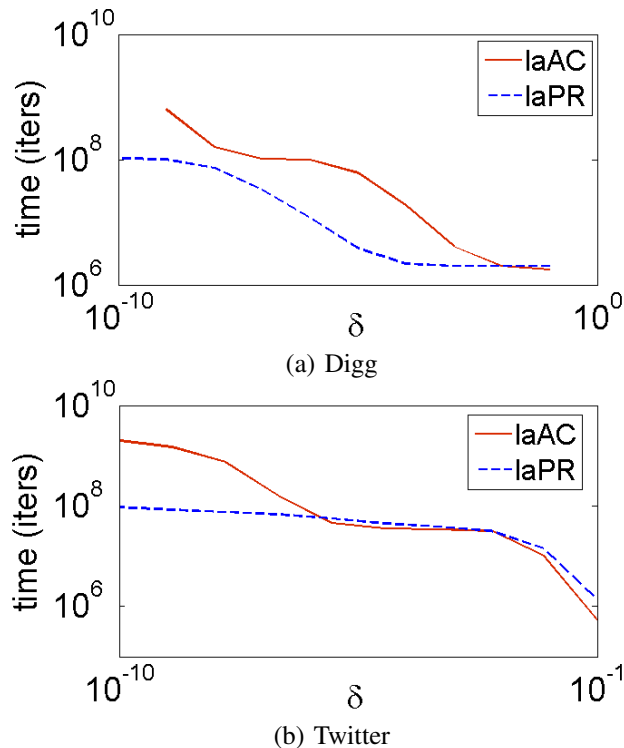


Fig. 5. Number of iterations required to compute limited attention centralities for different value of δ for (a) Digg and (b) Twitter networks.

Everyone's an influencer: quantifying influence on twitter. In *Proc. 4th ACM Int. Conf. on Web search and data mining*, pages 65–74, 2011.

- [3] A. Bavelas. A mathematical model for group structures. *Human Organization*, 7:16–30, 1948.
- [4] P. Bonacich. Power and centrality: a family of measures. *Am. J. Sociology*, 92(5):1170–1182, 1987.
- [5] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201, 2001.
- [6] S. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, January 2005.
- [7] R. S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA, 1995.
- [8] R. S. Burt. Structural holes and good ideas. *The American J. Sociology*, 110(2):349–399, 2004.
- [9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc. 4th Int. Conf. on Weblogs and Social Media (ICWSM)*, 2010.
- [10] K. M. Frahm and D. L. Shepelyansky. Google matrix of twitter, 2012.
- [11] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.
- [12] R. Ghosh and K. Lerman. Predicting Influential Users in Online Social Networks. In *Proc. KDD workshop on Social Network Analysis (SNAKDD)*, May 2010.
- [13] R. Ghosh and K. Lerman. Parameterized centrality metric for network analysis. *Physical Review E*, 83(6):066118+, June 2011.
- [14] R. Ghosh and K. Lerman. Rethinking centrality: The role of dynamical processes in social network analysis. *submitted to J. Discrete and Continuous Dynamical Systems*, 2012.
- [15] R. Ghosh, K. Lerman, T. Surachawala, K. Voevodski, and S.-H. Teng. Non-Conservative diffusion and its application to social network analysis. Technical report, University of Southern California, Feb 2011.
- [16] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proc. 27th Int. Conf. on Human factors in computing systems*, pages 211–220, 2009.
- [17] Michael Goldhaber. The Attention Economy and the Net. *First Monday*, 2(4-7), 1997.
- [18] M. S. Granovetter. The Strength of Weak Ties. *American J. Sociology*, 78(6):1360–1380, 1973.
- [19] D. Gruhl, R. Guha, D. L. Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. 13th Int. Conf. on World Wide Web*, pages 491–501, 2004.
- [20] N. O. Hodas and K. Lerman. How limited visibility and divided attention constrain social contagion. In *ASE/IEEE Int. Conf. on Social Computing*, 2012.
- [21] B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *J. Information Science*, 35(6):758–765, December 2009.
- [22] D. Kahneman. *Attention and effort*. Prentice Hall, 1973.
- [23] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.
- [24] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network, 2003.
- [25] D. Krackhardt and J. R. Hanson. Informal networks: the company behind the chart. *Harvard business review*, 71(4):104–111, 1993.
- [26] R. Lambiotte, R. Sinatra, J. C. Delvenne, T. S. Evans, M. Barahona, and V. Latora. Flow graphs: Interweaving dynamics and structure. *Physical Review E*, 84(1):017102+, July 2011.
- [27] C. Lee, H. Kwak, H. Park, and S. Moon. Finding Influentials from Temporal Order of Information Adoption in Twitter. In *Proc. 19th World-Wide Web (WWW) Conf. (Poster)*, 2010.
- [28] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proc. 13th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, KDD '07*, pages 420–429, 2007.
- [29] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.
- [30] R.A. Rensink, J.K. O'Regan, and J.J. Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8(5):368, 1997.
- [31] X. Tang and C. C. Yang. Ranking user influence in healthcare social media. *ACM Trans. Intell. Syst. Technol.*, 3(4), September 2012.
- [32] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2, March 2012.
- [33] F. Wu and B. A. Huberman. Novelty and collective attention. *Proc. Nat. Academy Sciences USA*, 104(45):17599–17601, 2007.